DOCUMENT RESUME

ED 236 208 TM 830 719

AUTHOR Francis, Alexandria S.; Holmes, Susan E.
TITLE Criterion-Referenced Standard-Setting in Certification and Licensure: Defining the Minimally Competent Candidate.
PUB DATE Aug 83
NOTE 20p.; Paper presented at the Annual Meeting of the American Psychological Association (Anaheim, CA, August 26-30, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Information Analyses (070)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Certification; *Criterion Referenced Tests; *Cutting Scores; *Interrater Reliability; Measurement Techniques; *Minimum Competencies; Standards; Testing Problems
IDENTIFIERS *Standard Setting

ABSTRACT

Discrepancies among the standards produced by different criterion-referenced standard-setting techniques may be the result of a failure to adequately define the minimally competent candidate. Current research in this area is reviewed in terms of three categories: studies in which no formal assistance in conceptualization is given to judges, studies in which conceptualization is accomplished through discussion including the review of relevant materials, and studies in which empirical information about test items or test performance is used. Two standard-setting studies using a combination of these techniques are discussed in detail. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED236208

Criterion-Referenced Standard-Setting in Certification and Licensure: Defining the Minimally Competent Candidate

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

A. S. Francis

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Alexandria S. Francis, Ph.D.

and

Susan E. Holmes, Ph.D.

CTB/McGraw-Hill

Monterey, California

Prepared for presentation at the

Annual Meeting of the American Psychological Association

Anaheim, California

August 26-30, 1983

TM 830 715

Criterion-Referenced Standard-Setting in Certification and Licensure: Defining the Minimally Competent Candidate

Alexandria S. Francis, Ph.D.

Susan Holmes, Ph.D.

CTB/McGraw-Hill

Monterey, California

Discrepancies among the standards produced by different criterion-referenced standard-setting techniques may be the result of a failure to adequately define the minimally competent candidate. Current research in this area is reviewed in terms of three categories: studies in which no formal assistance in conceptualization is given to judges, studies in which conceptualization is accomplished through discussion including the review of relevant materials, and studies in which empirical information about test items or test performance is used. Two standard-setting studies using a combination of these techniques are discussed in detail.

Criterion-Referenced Standard-Setting in Certification and Licensure: Defining the Minimally Competent Candidate

Alexandria S. Francis, Ph.D.

Susan E. Holmes, Ph.D.

CTB/McGraw-Hill

Monterey, California

Introduction

Professional and occupational testing for certification and licensure has increased substantially over the last ten years. The increased use of tests for this purpose is due in part to the increased regulation of professions and occupations by state governments. However, many other occupations require certification by nongovernmental agencies. In addition, requirements for relicensure and recertification are becoming more common, sometimes involving reexamination.

Although the exact figures are not known, a 1976 survey by Nafziger and Hiscox estimated that "over 2000 occupations have licensing or certification procedures of some sort." The range of professions and occupations currently requiring licensure or certification is quite wide. Included are traditional professions such as medicine, psychology, law, and education. Also included are a wide variety of technical and specialized occupations such as air traffic controller, barber, electrician, laboratory technician, mechanic, landscape architect, and real estate agent.

The primary purpose of certification and licensure is to protect the public from incompetent practitioners. Licensure requirements are screening

standards to ensure that licensed individuals will possess enough knowledge and skill to perform important occupational activities safely and effectively. Within this context, defining the minimum level of competence required to protect the public becomes one of the most important, and difficult, tasks in most licensure programs. It is usually defined in terms of a required level of education, experience, and/or a test score. The focus of this paper is on the latter, referred to variously as a "cut-off score" or a "passing score." The process by which such a score is determined is called standard-setting.

Standard-setting is a critical component of every licensure examination program. Because of state and federal regulations like the Uniform Guidelines, New York's and California's Truth-in-Testing laws, and the Gibbons Bill, the current trend is toward the use of a criterion-referenced approach to the standard-setting process. Here candidate test performance is compared to some absolute standard or criterion judged to represent the minimum level of competence acceptable for licensure, rather than to the performance of other examinees tested. The latter is referred to as a norm-referenced approach.

The issue in licensure is whether the candidate is safe and effective, and not whether he or she is safe and effective compared to other candidates. For this reason, the more traditional norm-referenced approach is being seriously questioned. Safe and effective practice has an absolute, rather than relative, nature.

A variety of criterion-referenced standard-setting procedures are currently available. Some involve judgments about test content or test questions,

while others involve judgments about individuals or groups of individuals. Three popular methods based on judgments about test content or test questions are the Angoff method (1971), the Nedelsky method (1954), and the Ebel method (1972). Each consists of five basic steps:

1. Select the judges.
2. Define the minimally competent candidate.
3. Train the judges in the method to be used.
4. Make the judgments.
5. Combine the judgments to obtain a passing score.

The first two steps are the same for all three methods. The remaining steps differ with the method employed.

All of the methods are based on the concept of the "borderline" candidate. Within the context of certification and licensure, such a candidate would be minimally competent, possessing just enough knowledge and skill to be competent to practice. In each method, judgments about how such a candidate is likely to respond are made for each question in the examination. The kind of judgment required, however, depends on the particular method being used. In all cases, the methods result in a passing score on the examination below which minimal competence has not been demonstrated.

Several reviews of the rather extensive literature on standard-setting are currently available (Glass, 1978; Hambleton, 1980; Jaeger, 1979; Shepard, 1980). For a detailed discussion of each of the methods outlined above, see Livingston and Zieky (1982).

Despite similarities among the three popular criterion-referenced standard-setting methods, quite <u>different</u> results have been reported in studies where different methods were applied to the same examination (Brennan

& Lockwood, 1980; Koffler, 1980; Poggio, Glasnapp & Eros, 1982; Saunders, Ryan & Huynh, 1981; Skakun & Kling, 1980). Although disconcerting to the practitioner, such inconsistencies are not surprising. The methods themselves are based on quite different assumptions about the test behavior of the minimally competent candidate. For example, the Angoff method supposes that the candidate responds independently to items and that he or she has different probabilities of success for different items. On the other hand, the Nedelsky method asserts that the mimimally competent candidate answers test items by eliminating answer alternatives and then choosing at random from among the remaining ones. To expect methods with such different underlying assumptions to produce similar passing scores seems unreasonable.

It has been argued that these standard-setting methods should be avoided entirely because of the inconsistencies noted above (Glass, 1978). However, this position seems too extreme. It is the opinion of these authors that such technqiues can lead to usable standards <u>if</u> the choice among them is made consciously and includes a consideration of the assumptions underlying them.

More disconcerting, however, are those studies in which several applications of the <u>same</u> standard-setting method produced <u>different</u> passing scores (Andrew & Hecht, 1976; Halpin, Sigmon & Halpin, 1983; Mills & Barr, 1983; Saunders, Ryan & Huynh, 1981). In this case, the source of inconsistency can not be differences among methods because only one method is employed. Differences among passing scores developed using the same method must therefore be introduced into the standard-setting process in other ways. Several potential sources of these inconsistencies will be described next.

One potential source of error in the standard-setting process is specific attributes or characteristics possessed by the panel of judges performing the ratings. This is a particular concern when the number of judges on the panel is small. Several studies have examined the relationship between various characteristics of the judges and the standards they set (Behuniak, Archabault, & Gable, 1982; Halpin, Sigmon, & Halpin, 1983; Saunders, Ryan, & Huynh, 1981). Some of the characteristics considered included years of work experience, educational level, position, and grade level taught. Although inconclusive, the evidence suggests that such attributes do represent a potential source of error. Careful consideration should, therefore, be given to the selecting of judges, both in terms of their numbers and the representation of important characteristics.

A second potential source of error has been referred to as "intrajudge inconsistency" (Van der Linden, 1982). This kind of inconsistency arises when judges misperceive the way an item will actually function. As a result, ratings are obtained that are inconsistent with the actual performance of the items. Van der Linden (1982) has proposed a method for analyzing intrajudge inconsistency using item response theory in an attempt to minimize this problem.

A third potential source of error is defining the minimally competent candidate. This is a very critical component of the standard-setting process. There is simply no reason to expect judges to produce consistent standards unless they possess the same concept of minimally competent performance. This source of inconsistency has been described as "interjudge inconsistency" (Van der Linden, 1982), and it is the result of employing different

interpretations of minimally acceptable competence.

The focus of this paper is on the process of conceptualizing the minimally competent candidate in criterion-referenced standard-setting. Little attention has been paid to this potentially dangerous source of error. In an attempt to bring to light some of the issues involved, the results of a recent review of the criterion-referenced standard-setting literature are reported. The studies reviewed are categorized according to how conceptualization of the minimally competent candidate was accomplished. Specific examples of studies in each category are given. A detailed description of the procedure employed in two standard-setting studies is then presented.

Review of Current Research

A review of the current research on criterion-referenced standard-setting indicated that little information is generally provided about the process used to establish a definition of the minimally competent candidate. The studies reviewed generally fell into one of three major categories. They are as follows:

1. Studies in which no formal assistance in defining the minimally competent candidate was given to judges. Studies within this category assumed that judges already had a consistent definition of the minimally competent candidate and that no training or assistance was, therefore, required (Andrew & Hecht, 1976; Halpin, Sigmon & Halpin, 1983; Harasym, 1981; Mills, 1982).

2. Studies in which conceptualization was accomplished through a review of materials and discussion. Studies within this category reported a variety of techniques to assist judges in defining the minimally competent candidate. Some of the techniques used included the following:

   a. Sending materials to judges for review prior to the actual standard-setting session (Poggio, Glasnapp, & Eros, 1982). Examples of the materials sent were an overview on standard-setting and a dicusstion of related issues, a description of the specific standard-setting method to be employed, a test content outline, a list of required competencies, and a history of the development of the

examination (Ward, 1982) being reviewed. During the standard-setting session, additional information, as well as predetermined definitions of the minimally competent candidate, were sometimes provided to the judges (Behuniak, Archabault & Gable, 1982). For example, Saunders, Ryan, and Huynh (1981) reported that "minimum acceptable performance was defined as the lowest level of performance on the test for which a grade of 'B' would be awarded."

b. Using structured as well as unstructured discussions among judges during the standard-setting session. Structured discussions included the presentation of examples of competent, incompetent, and minimally competent behavior in a variety of job-related situations. The judges discussed these situations and then described the same behaviors for additional situations that were provided (Holmes, Francis & Ference, 1982a, 1982b). Unstructured, open-ended discussions allowed the judges to think about candidates in their own programs who were minimally competent, to share their personal experiences with such candidates, and to use the performance of these candidates as a basis for their ratings (Koffler, 1980; Skakun & Kling, 1980).

c. Rating sample items similar to the items on the examination to be reviewed. Judges rated sample items individually. They were then asked to share their ratings with the group and to explain the reasons for their choices (Ward & Marshall, 1982).

d. Reconciling large discrepancies in item ratings. During the actual rating process, items which produced ratings that differed beyond some prespecified amount were noted. Periodically throughout the session, these items were presented to the judges for a reconsideration. The judges discussed their initial ratings and in some instances were given the opportunity to change these ratings after the discussion. This technique, while generally producing no significant change in the final passing score (Brennan & Lockwood, 1980), does provide judges with the opportunity to continually evaluate their concept of the minimally competent candidate.

3. Studies in which empirical information about items or test performance was used in the conceptualization process. Although standard-setting is usually done prior to the administration of an examination, in some instances empirical information about the items on the examination being reviewed may be available at the time of the standard-setting session. Some studies reported using such infomation to facilitate the conceptualization process. Some of the technqiues in which empirical information was used to define the minimally competent candidate included:

a. Using item difficulties (p-values) obtained from some previous administration of the items to help judges develop a sense of overall group performance. In some cases, this kind of information was made available for sample items used during the conceptualization process. In other instances, it was actually incorporated into the item rating process (Mills & Barr, 1983).

b. Using empirical information about the test being reviewed. Some studies provided the judges with information about the level and the range of difficulty of items on the examination being reviewed. Information was sometimes provided about the characteristics of score distributions (e.g., skewness and changes in slope) for selected ethnic groups and programs (Ward, 1982).

An Application

While techniques for defining the minimally competent candidate can be categorized as above, many studies reviewed reported using a combination of one or more of these techniques in the conceptualization process. Such an approach was used by Holmes, Francis, and Ference (1982a, 1982b) when defining minimally competent, entry-level nursing candidates. The following is a description of the techniques used in two criterion-referenced standard-setting studies conducted for the national licensure examinations for registered and practical nurses.

Initially, a variety of materials were sent to judges several weeks before the standard-setting session. Included in the packet each judge received was a test content outline, documents describing criterion-referenced testing and the Modified Angoff Method, a list of critical nursing behaviors, and an agenda for the standard-setting session. In a cover letter accompanying each packet, the judges were asked to carefully study all the materials in their packets in preparation for the standard-setting session.

At the standard-setting session, the judges were presented with the objectives for the session. They were given detailed instructions on the standard-setting technique to be employed, and they received information about the examination to be reviewed and the population of candidates who sit for it. They were then asked to review and discuss several job-related situations based on the test plan that had been developed earlier. Examples of competent, incompetent, and minimally competent behaviors were discussed for

each situation. Additional job-related situations were then presented and the judges were asked to develop their own examples of competent, incompetent, and minimally competent behaviors for each.

After a consensus had been reached about what constituted minimally competent behavior in each of the situations presented, the judges were given 25 sample questions from a previous form of the examination. They were asked to determine the percentage of minimally competent candidates who would be able to correctly answer each item. This was exactly the kind of judgment required by the Angoff method. Upon completion, the judges shared their ratings of the sample items, and the item ratings were averaged for each item. The judges were then given a table that contained the actual percentage of the candidates who answered each sample item correctly. These values were compared to those obtained after averaging the judges' ratings. Any discrepancies were noted and discussed.

Also included in the table were percentages for a hypothetical group of competent, incompentent, and minimally competent candidates. Minimally competent candidates here were arbitrarily defined as those candidates whose scores were at or above the passing point of the previous test by no more than half the test's standard error of measurement. The percentages for all three groups were compared to those obtained after averaging the judges' ratings, and again discrepancies were noted and discussed. Further discussion among the judges was encouraged until they reached consensus on the definition of the minimally competent candidate.

The judges then independently rated all the items on the actual examination. Periodically throughout the rating process, they were asked to reconsider items for which large discrepancies in rating were obtained. These items were discussed, and the judges were given the opportunity to change their ratings if they so desired. The final ratings of the judges were then averaged for each item, converted to a proportion, and summed across the entire examination to obtain a passing score.

Conclusions

In 1978, a symposium on the legal aspects of credentialing was jointly sponsored by the American Educational Research Association and the National Council on Measurement in Education. The participants of that symposium concluded that standard-setting was the most challenging issue facing measurement specialists at that time. This conclusion continues to be true today, particularly in the area of certification and licensure testing.

The focus of this paper was on one potentially dangerous source of error in the standard-setting process, namely the conceptualization of the minimally competent candidate. Failure to ensure that all judges possess a consistent definition of the minimally competent candidate precludes the possibility of obtaining a consistent standard. A review of the research on criterion-referenced standard-setting indicated that little information has generally been reported about how conceptualization is accomplished, despite the obvious implications of interjudge inconsistencies. Studies that did provide information about the conceptualization process generally fell into three categories:

1. Studies in which no formal assistance in defining the minimally competent candidate was given to judges;
2. Studies in which conceptualization was accomplished through a review of materials and discussion;
3. Studies in which empirical information about items or test performance was used in the conceptualization process.

Many of the studies reviewed reported using a combination of one or more of these techniques.

To date, no research has been reported on the effects of various techniques for conceptualizing the minimally competent candidate on the standards being established. Because of the political and social implications of determining a passing score in certification and licensure testing, studies examining these issues are strongly recommended.

References

Andrew, B.J., & Hecht, J.T. A preliminary investigation of two procedures for setting examination standards. Educational and Psychological Measurement, 1976, 36, 45-50.

Angoff, W.H. Scales, norms, and equivalent scores. In Thorndike (Ed.), Educational measurement (2nd ed.). Washington, DC: American Council on Education, 1971.

Behuniak, P., Jr., Archabault, F.X., Jr., & Gable, R.K. Angoff and Nedelsky standard setting procedures: Implications for the validity of proficiency test score interpretation. Educational and Psychological Measurement, 1982, 42, 247-255.

Brennan, R.L., & Lockwood, R.E. A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. Applied Psychological Measurement, 1980, 4, 219-240.

Ebel, R.L. Essentials of educational measurement. Englewood Cliffs, NJ: Prentice-Hall, 1972.

Glass, G.V. Standards and criteria. Journal of Educational Measurement, 1978, 15, 237-261.

Halpin, G., Sigmon, G., & Halpin G. Minimum competency standards set by three divergent groups of raters using three judgmental procedures: Implications for validity. Educational and Psychological Measurement, 1983, 43, 185-196.

Hambleton, R.K. Test score validity and standard-setting methods. In R. A. Berk (Ed.), Criterion-referenced measurement: The state of the art. Baltimore, MD: John Hopkins University Press, 1980.

Harasym, P.H. A comparison of the Nedelsky and modified Angoff standard-setting procedure on evaluation outcome. Educational and Psychological Measurement, 1981, 41, 725-734.

Holmes, S.E., Francis, A.S., & Ference, H.M. Modified Angoff passing score workshop (registered nurse), final report. Report prepared by CTB/McGraw-Hill for the National Council of State Boards of Nursing, June 1982. (a)

Holmes, S.E., Francis, A.S., & Ference, H.M. Modified Angoff passing score workshop (registered nurse), final report. Report prepared by CTB/McGraw-Hill for the National Council of State Boards of Nursing, October 1982. (b)

Jaeger, R.M. Measurement consequences of selected standard-setting models. In M.A. Bunda and J.R. Saunders (Ed.), Practices and Problems in Competency-Based Measurement. National Council on Measurement in Education Monograph, Washington, D.C., 1979.

Koffler, S.L. A comparison of approaches for setting proficiency standards. Journal of Educational Measurement, 1980, 17, 167-178.

Livingston, S.A., & Zieky, M.J. Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service, 1982.

Mills, C.N. A comparison of three methods of establishing cut-off scores on criterion-referenced tests. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, March 1982.

Mills, C.N., & Barr, J.E. A comparison of standard setting methods: Do the same judges establish the same standards with different methods? Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, April 1983.

Nafziger, D.G., & Hiscox, N.D. A survey of occupational licensing and certification procedures. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, California, April 1976.

Nedelsky, L. Absolute grading standards for objective tests. Educational and Psychological Measurement, 1954, 14, 3-19.

Poggio, J.P., Glasnapp, D.R., & Eros, D.S. An evaluation of contrasting-groups methods for setting standards. Paper presented at the Annual Meeting of the American Educational Research Association, New York, March 1982.

Saunders, J.C., Ryan, J.P., & Huynh, H. A comparison of two approaches to setting passing scores based on the Nedelsky procedure. Applied Psychological Measurement, 1981, 5, 209-217.

Shepard, L. Standard setting issues and methods. Applied Psychological Measurement, 1980, 4, 447-467.

Skakun, E.N., & Kling, S. Comparability of methods for setting standards. Journal of Educational Measurement, 1980, 17, 229-235.

Van der Linden, W.J. A latent trait method for determining intrajudge inconsistency in the Angoff and Nedelsky techniques of standard setting. Journal of Educational Measurement, 1982, 19, 295-308.

Ward, A.W. Using a Rasch calibrated item pool in setting and maintaining passing score standards. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, March 1982.

Ward, W.H., & Marshall, D.M. The ability of judges to subjectively estimate item difficulty. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, March 1982.